图书情报工作
第 65 卷 第 1 期 2021 年 1 月

科学论文语义增强的研究进展与趋势研判

■ 宋宁远[1] 裴雷[1] 王春迎[2]
[1] 南京大学信息管理学院 南京 210023 [2] 郑州大学信息管理学院 郑州 450001

摘 要：[目的/意义]随着科学交流体系向电子媒介迁移，传统的科学论文内容组织及呈现方式带来了诸多弊端。科学论文语义增强能够创新科学论文内容的组织与呈现方式，是解决这些问题的关键，得到了来自科研机构与学术出版商的重视，形成了一系列理论与实践成果。对这些成果进行梳理、归纳，发现其中的优势与不足，能够为后续推动科学论文语义增强的进一步发展起到指导作用。[方法/过程]从语义增强的概念入手，着重分析科学论文语义增强的核心目标、实现路径与关键问题，随后，梳理对科学论文中正文本与副文本内容进行语义增强的理论与实践成果，并围绕科学论文语义增强路径上的三个阶段：语义标注、语义组织与可视化呈现进行对比分析。[结果/结论]研究进一步归纳总结现阶段科学论文语义增强的特点，并对科学论文语义增强的未来发展及研究提出 4 点意见。

关键词：科学论文 语义增强 语义标注 语义组织 可视化
分类号：G255
DOI：10.13266/j.issn.0252-3116.2021.01.013

1 引言

科学论文是科学交流的载体形式，具有特定的结构与形式。随着期刊数字化转型，电子期刊已经成为主流。在论文发表量日益增大和单篇文章平均阅读时间减少的情况下，利用新兴的数字技术对科学论文内容组织与呈现方式进行创新，进而提升读者的阅读与交流效率，逐渐引起学界重视[1-2]。借助本体、RDF、关联数据等语义技术，在准确表征科学论文内容语义功能的基础上，实现细粒度知识片段的可视化关联与发布，不仅有助于提升读者对科学论文的内容的阅读与理解，也有利于计算机处理和挖掘学术文本，实现基于数据的自动知识发现[3]。这一系列技术与方法开启了科学论文的语义增强（Semantic Enhancement 或 Enrichment）之路。

语义增强是针对内容进行的增值性编辑加工活动，可以提高数字内容资产的价值。目前，对科学论文进行语义增强已经受到学界与业界的广泛重视。D. Shotton 等曾进行了一系列语义出版实验[4]，探索了对论文进行语义增强的方法和路径。英国皇家化学学会[5]、爱思唯尔公司[6]、*Nature* 杂志，以及微软、Google 和部分文化遗产机构也开展了学术出版和网络资源的语义增强实验。借助语义增强，创新科学论文内容组织方式及表现形式，可以提高科学论文资源的利用效率，挖掘科学论文内容的潜在价值，实现论文内容与知识的互联互通，促进科技情报工作向智慧服务转型升级。

经过数十年的尝试及探索，针对科学论文的语义增强取得了众多研究及实践成果，对现有研究及实践成果进行系统梳理，不仅有助于进一步明确科学论文语义增强的方式方法及其实现路径，更有助于明确未来语义增强研究的发展方向、重点及趋势，将会对科学论文内容与价值的再利用提供方向性指引。

为了更好地对科学论文语义增强的进展进行综述与对比，需要进行文献检索。本研究首先以科学论文、语义增强为关键词在中国知网中进行检索，发现相关文献较少，随后使用本体、语义标注、语义组织等为关键词进行检索，共返回中文文献记录 142 条。其次，本研究使用 scientific papers，semantic enrichment，semantic enhancement，ontology，semantic annotation，semantic

作者简介：宋宁远（ORCID:0000-0001-5601-1487），博士后；裴雷（ORCID:0000-0003-4754-4112），教授，博士，博士生导师，通讯作者，E-mail:plei@nju.edu.cn；王春迎（ORCID: 0000-0003-4767-4523），讲师，博士。
收稿日期：2020-01-18 修回日期：2021-01-19 本文起止页码：82-90 本文责任编辑：王传清

organization 等关键词在 Web of Science 核心数据集中进行检索,共返回外文文献记录 271 条。通过人工排除相关性较弱的文献,共获取 322 篇文献,作为综述及对比分析的样本。同时,为了更为全面地归纳语义增强的路径,研究还特别调查了施普林格·自然、威利、爱思唯尔等大型出版机构的语义增强项目,一起作为综述的对象进行分析。

2 科学论文语义增强内涵、需求及实现方式

2.1 科学论文语义增强概念发展

语义增强是伴随计算机文本处理技术与语义网的发展而在信息资源管理与科学论文出版领域兴起的新概念,旨在解决现有的电子文档语义揭示和编码表示不足的问题。目前,科学论文多以 HTML 和 PDF 格式文档为主。囿于文档编码方案的不足,这两类文档普遍缺乏语义标记,由此导致网络搜索引擎难于理解这些文档中的内容片段和元素的语义特征及功能。因此,读者也难于检索和利用文档的细粒度片段及知识点,所以对文档进行语义增强成为数据资源改造升级和语义网建设不可回避的环节。

在语义增强的内涵理解上,V. Damjanovic 认为语义增强与语义检索、语义组织、语义标注及语义分析与知识发现等多种活动有关[7]。Europeana 在语义增强框架中,定义了语义增强的基本环节,包括分析、关联与增强[8]。国际图联的 LRM 模型定义了语义增强环境下用户的 5 种信息任务,包括找寻、识别、选择、掌握与探索[9]。SURF 基金会报告指出科学论文的语义增强是就要集成研究数据、辅助材料、数据记录、公开发表的出版物等为主要手段,实现对传统论文内容的延伸与扩展[10]。M. Hoogerwerf 认为科学论文语义增强是以对象为基础的信息集成,对象泛指各种多媒体要素和文本块,如视频、用户评论以及数据库等[11],这些对象之间存在显著的关联。L. Breure 等认为科学论文语义增强应当在完备的语义元数据体系下,支持线性和非线性阅读[12]。

综合以上不同研究针对语义增强的观点、实施阶段及侧重点的理解,本文认为针对科学论文的语义增强是以提升用户阅读效率与知识获取效果为主要目的,综合利用多种语义技术与可视化技术,对科学论文进行一系列结构化、语义化、关联化、可视化处理。语义增强的主要阶段包括语义标注、语义关联与可视化呈现。

2.2 科学论文语义增强核心目标

对科学论文进行语义增强主要是为了充分揭示蕴含在科学论文内部的潜在知识,创新科学论文内容组织与呈现方式,提高用户的阅读效率与阅读效果。即通过语义增强构建具备可信的、情境化的、关联的、可认知、可预测、可利用的智慧数据集,实现由传统文献资源到智慧数据的转换与升级,以充分挖掘蕴含在科学论文内容中的潜在知识,并在内容数据充分关联的基础上,借助可视化技术提高用户获取信息的效率与效果。

围绕核心目标,科学论文语义增强具备多种应用场景:知识发现、语义出版与策略型阅读。在知识发现领域,富语义的科学论文内容数据为从不同视角分析科学论文提供可能,实现知识抽取、知识检索、知识发现等高级应用。在语义出版领域,借助语义增强,可以实现出版对象由篇章层次的科学论文向细粒度陈述的过渡。在策略型阅读领域,通过不同粒度科学论文内容语义特征的揭示,定位对用户最有价值的信息。

2.3 科学论文语义增强实现方式与关键问题

2.3.1 实现方式

科学论文是一种复杂的知识系统,由大量的正文本与副文本内容组成,副文本内容主要包括题录信息、摘要、引用及参考文献信息;正文本内容是指蕴含了大量知识的科学论文内容。其中,副文本内容的主要作用是用来辅助理解正文本,并对正文本进行解释说明。科学论文语义增强是为了对科学论文正文本与副文本内容进行语义表征,创新科学论文内容组织及呈现方式,生成适用于提升用户阅读效果的增强型论文。因此,科学论文语义增强的实现路径一般包括:语义标注、语义组织与内容可视化,如图 1 所示:

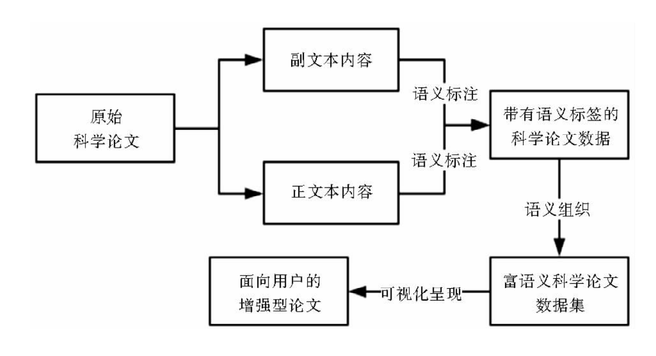


图 1 科学论文语义增强实现方式

(1)语义标注。语义标注是指将科学论文中的实体与本体、主题词表等知识组织工具中的概念进行关联,利用本体中定义的概念、属性与关系揭示科学论文的语义特征,对科学论文进行语义描述,并生成带语义

标签的语义内容(semantic content),由此实现机器可读。语义标注是实现科学论文语义增强由以文献为中心(document-centric)向以实体为中心(entity-centric)转变的重要过程[13]。

(2)语义组织。语义组织是在对科学论文进行语义标注的基础上,实现对所生成的带有语义标签的语义内容进行关联与组织。语义组织过程涉及组织模型的设计、本体互操作、本体映射等工作,融合了多种本体及元数据集。语义组织的结果是生成相互关联的富语义科学论文内容数据集。

(3)内容可视化。内容可视化是综合利用多种计算机视觉技术,对科学论文内容数据集进行图形化、多媒体化呈现,生成适用于用户的增强型论文,以提高内容的感知能力,进而促进用户的知识获取效率。

2.3.2 关键问题

在"语义标注 - 语义组织 - 内容可视化"的路径上,为了更好地实现科学论文语义增强,还需要着力解决以下关键问题。

(1)对正文本内容信息的多维语义描述。科学论文正文本内存在大量非结构化内容数据,对这些数据全面表征可揭示科学论文内容的组织模式与基本架构,能够实现从文献层次向细粒度内容层次的过渡,是促进科学论文语义增强进一步发展的关键。

(2)多源数据的语义关联、组织与发布。科学论文语义增强需要创新科学论文组织模式,这就需要对经过语义标注的多源数据(题录信息、引用信息及内容数据)进行充分地关联与组织,包括设计语义组织模型、选择组织与发布工具等。其中的关键在于对科学论文的逻辑结构、语义关系以及引用关系进行准确描述与规范定义,进而构建适用于不同应用场景的组织模型。

(3)语义内容数据的可视化呈现。内容数据的可视化、可交互呈现是提高用户内容理解效率的主要方式。除了对字、词、概念等进行可视化呈现之外,尤其需要考虑如何利用图形准确表征科学论文逻辑结构、论证方式、特定内容等富语义内容数据。

3 科学论文语义增强路径分析

下文将从科学论文副文本及正文本语义增强的不同方式进行综述。

3.1 副文本内容语义增强

3.1.1 题录信息语义关联与组织

科学论文的题录信息包括文章标题、作者信息、摘要、关键词、项目与基金信息等,其信息格式明确,可以通过诸如都柏林核心元数据集等进行描述。现阶段,对题录信息进行语义增强的主要方式是通过设计书目本体实现对题录信息的语义描述,并通过多本体的协同使用,实现科学家、论文、会议、期刊等多源信息的语义关联。

对题录信息的语义描述以书目信息本体(the Bibliographic Ontology Specification, BIBO)与FRBR对应书目信息本体(FRBR-aligned Bibliographic Ontology, FaBiO)为代表。BIBO共定义了69个元素,其中最主要的是对文献类型的定义[14]。FaBiO在BIBO的基础上,融合了FRBR框架中关于作品、内容表达、载体表现和单件的分类,同时也包括了对创作者和创作团体描述,最终形成了整合性的本体[15]。在书目本体对单篇文献题录信息语义描述基础上,VIVO本体系统[16]整合了BIBO、FOAF、DC等本体与元数据集,添加大量语义关系,建构了科学家的信息交流语义模型。

3.1.2 摘要语义增强

摘要是对科学论文主要内容的归纳与总结,本身也蕴含着较为丰富的内容信息,因此针对摘要部分的增强方式也较为丰富。

喻琪琛等[17]总结了采用不同语义增强方式的摘要,结构化摘要主要通过对一段式摘要添加相应的语义元素(背景、目的、方法、结果、讨论等),以明晰摘要的结构、丰富内容的语义,便于用户快速掌握论文的重点内容;视频摘要与图形摘要利用图表、音视频与文字结合的方式,对摘要内容进行多媒体、可视化地表达;结构化数字摘要主要面向摘要内容的机器可理解性,并通过实体链接等实现与外部知识库的关联;亮点摘要则揭示了论文中最为重要的断言与陈述,具备较高的情报价值。

3.1.3 引用功能语义描述

引文与参考文献信息通常包括被引文献的作者、论文标题、期刊、出版商等,此外,引文信息关联了被引文献与施引文献,构成了引用关系,内涵了一定的语义属性,诸如引用情感、引用情境等,同样是语义增强需要关注的重点。

现阶段,对引文与参考文献的语义增强主要通过构建相关本体,具有代表性的本体包括:引文类型本体(Citation Typing Ontology, CiTO)[18]与引用数量及引用环境本体(Citation Countingand Context Characterization Ontology,C4O)[19]。CiTO借助RDF表示引用关系,并对其语义属性进行定义。在CiTO中,引文语义主要由

修辞关系及事实关系两方面进行定义,修辞关系主要指作者的引用情感,包括积极、中性、消极三类;事实关系则体现引文的作用,包括引用数据、引用方法等。C4O 主要用来对同一参考文献在不同文献中的引用位置、引文环境进行定义,同时也与谷歌学术等相关联,实现对总体引用次数的描述。

CiTO、C4O 等均具有较强的扩展性,可以同 FOAF 本体、都柏林核心元数据集进行关联,用以表示引用文献的作者信息。同时也可以与篇章元素本体(Discourse Element Ontology, DEO)、文献组件本体(Document Component Ontology, DoCO)等出版物内容本体进行较好的协同,用以实现对细粒度引用情境的表征。

3.1.4 副文本内容的关联与发布

在语义组织与关联发布方面,副文本内容的语义增强也较为成熟,形成了一定数量与规模的开放数据集与知识图谱。

在数据集建设与发布方面,OpenCitations 数据集最具代表性,它是通过众包形式建设的论文结构化信息数据集,主要包括了会议论文、图书章节、期刊论文等的题录信息,使用 CiTO、FaBiO 等本体进行语义描述[20],并使用 RDF 语言,将经 CrossRef 和 ORCID 标引的文献数据进行关联开放。

知识图谱是较多出版及科研机构采用的对题录信息及参考文献信息进行关联与发布的形式。施普林格·自然在 2015 年启动了 SciGraph[21]项目,在知识组织的基础上,通过数据融合、知识发现、内容计算来实现多源异构数据的跨模态语义聚合。清华大学的 AMiner[22]科学知识图谱通过对科技文献、专家学者、学术活动等科技大数据进行分析挖掘,提供面向科技文献、专家学者和学术活动的语义搜索、语义分析、成果评价等知识服务。微软学术图谱(Microsoft Academic Graph, MAG)[23]通过智能分析网页学术实体及它们之间的关系所构建的异构知识图谱。此外,MAG 和 AMiner 合作构建了开放学术图谱(Open Academic Graph, OAG)[24],实现近 6 500 万对链接关系,可以支撑学术界对学者合作关系、学术主题挖掘等领域的研究。上海交通大学发布的学术知识图谱 AceKG[25],包含超过 1 亿个学术实体、22 亿条三元组信息,为每个实体提供了丰富的属性信息,旨在支持多元学术大数据挖掘项目。

3.2 正文本内容的语义增强

正文本内容包含有以字、词、词组为主的概念实体;以句子为表现形式的陈述、命题;由若干语句构成的内容组件(Component);以及借由组件之间关系(Relationship)而形成的逻辑结构。当前研究针对不同层次、不同粒度的内容进行了语义增强理论与实践探索。

3.2.1 科学论文概念实体的抽取与表示

在科学论文概念实体抽取方面,综合运用领域本体以及包括命名实体识别在内的自然语言处理技术,实现了对概念实体的抽取与语义表示。诸如微观概念地图的挖掘与构建[26]、学术概念属性的抽取[27]、关键术语抽取[28]等。在概念实体的可视化表示方面,现有研究与实践主要通过标签云、标签树等形式呈现科学论文的核心概念。

3.2.2 科学论文陈述的描述与关联发布

科学陈述是组成科学论文内容的基础,也是概念实体存在状态及属性的直接表现。当前最具有代表性的陈述表示模型即为纳米出版物(Nanopublication)。

纳米出版物是以“科学陈述”为单位的“具有科学意义、机器可读的、最小的可出版单元”模型[29]。该模型包含了核心科学陈述和相关语境,方便科学声明层面的知识处理工作,诸如科学声明的整合、查询、推理等。概括来说,纳米出版物主要由内容性和功能性组成部分构成。其中,内容性组成部分以概念三元组为基础,将每一个具有实际意义的三元组视为一条科学陈述。科学陈述与其出处信息构成了一条最基本的纳米出版物。除此之外,出版物信息(包括归属、整合时间、引用情况等)、支持性信息等则对纳米出版物起到了附加解释作用。目前,纳米出版物模型在生物医学及数字人文项目中得到了较为广泛的运用,形成了一定规模的纳米出版物数据集。

3.2.3 科学论文内容组件及逻辑结构语义表征

除了对以语句为基本单位的陈述进行语义描述与增强之外,也有研究从语篇分析的角度入手,提出了科学论文内容组件的概念。按照解读视角的不同,主要集中在以下 4 个方向:

(1)修辞与功能组件。围绕科学调查过程,在科学实验本体(EXPO)及 CISP 的基础上[30],M. Liakata 提出了 CoreSC(Core Scientific Concept)模型[31],科学文本中的陈述按照科学实验的不同过程划分为:假设、动机、目的、目标、背景、方法、实验、模型、观察、结果和结论。该模型详细定义了科学实验的过程,但是对于大量论述性文本的语义表征能力不足。

A. De Waard 认为科学论文是围绕具体科学目标而进行的知识建构,其在 2006 年提出 ABCDE 模型[32]。该模型从标注、背景、贡献、讨论及实体等 5 个部分对科学论文进行描述,不仅描述了文献内容(背

景、贡献、讨论),也定义了文献元数据(标注)及实体层面的信息(实体),但模型粒度较粗,表达能力有限。通过聚焦文献内容组成模块,其又提出了框架篇章块(Discourse Segment)模型,更细粒度地揭示了科学论文中的知识单元,该模型包括:事实、假设、目标、方法、结果、影响和问题等 7 个类别[33]。

L. Zhang 针对用户在科学论文阅读过程中产生的功能性需求,定义了学习背景知识、参考事实、参考论点、参考方法、跟进前沿研究等 6 种科学论文语境下的信息使用任务。此外,L. Zhang 等结合研究空间理论、体裁分析等,提出了包含 41 个功能单元的概念模型[34]。

(2)论证结构。对论证结构的语义描述一般包括对论证组件与论证关系的定义。在论证组件方面,S. Teufel 提出了论证块(Argumentative Zoning, AZ)模型[35]。论证块模型将科学文本中的不同的内容组件定义为目标、对比、基础、文本、背景等几个类别。随后,S. Teufel 对这一理论进行了拓展,将引用功能及作者的情感倾向与文本修辞功能进行结合,提出了更细粒度的框架 Argument Zoning II[36]。该框架定义了 14 种不同的修辞组件,新增加类型包括对比(CoDI)、指出缺陷、观点相悖、支持、使用等,尤其注重对不同观点的比较,更加适合科学论文的特点。N. L. Green 以生物医学领域的科学论文为例,研究了科学论文结构的表征问题[37],提出了包括假设、结论、背景知识等组成的论证框架用以表征科学论文的论证结构[38],并列举了使用该框架对论证结构进行表示的若干实例[39-40]。这些研究都较为清晰地定义了论证组件及其语义特征,但囿于论证关系定义的欠缺,在表征科学论文论证结构方面还存在不足。

在论证关系方面,较为成熟的项目是学术本体项目(Scholarly Ontologies Project)[41]。在该项目中,S. J. Buckingham Shum 等提出将科学论文分解成基本的篇章知识单元并基于认知关联关系等理论,实现了对论证关系的定义,分别包括:因果关系、问题相关关系、相似性关系、通用关系、支持/挑战关系、分类关系。每一类关系都包含了显式的极性(正面或负面),以及具体的权重。其研究结果催生了一系列对论证关系进行标注和可视化的工具[42-43]。

(3)情境信息语义描述。情境信息揭示了科学论文内容组件存在状态。P. Thompson 针对生物医学领域科学论文情境信息设计了 EventMine-MK 标引框架,使用知识类型、可信度等级、极性、来源、程度以及不同属性之间聚合而成的高维知识类型,对科学论文情境信息进行了表示[44]。此外,P. Thompson 等还设计了针对新闻事件的情境信息标注框架,该框架在级性、时间、体裁的基础上,增加了消息来源、语态、主观性等不同维度[45]。A. De Waard 等[46]提出了情境信息表示模型,包括确定性等级、基础和来源等三个维度。其中,确定性等级维度用来表示陈述的可信度情况;基础维度用来表示陈述、命题的存在状态;来源维度表征了陈述的出处信息。Claim Framework[47]由 C. Blake 提出,该框架以断言为主要描述对象,认为断言的组成要素除主体、客体等知识实体之外,还包括改变、方向、修饰、基础等情态要素。

以上三种表示模型给出了情境信息的不同定义,EventMine-MK 的使用对象为事件型知识,有利于事件知识表示与挖掘;A. De Waard 的模型更注重对陈述的多维表征;C. Blake 的框架建构了内容组件与概念实体间的存在关系,更侧重对实体之间逻辑关系的表征。

(4)科学论文内容本体。当前研究使用本体实现了对内容组件、组件间关系进行规范定义,设计并开发了大量科学论文内容本体。

粗粒度的修辞本体包括 SALT[48]、修辞块本体(Ontologyof Rhetorical Block, ORB)[49]等,宏观地定义了科学论文内容的修辞结构。细粒度的修辞本体以篇章元素本体[50]、文献组件本体[51]等为代表,细致地定义了科学论文的内容组件。除了修辞组件本体,Peroni 等提出了论证模型本体(Argument Model Ontology, AMO)[52],定义了包括断言、证据(evidence)、支撑、反驳、限定语、保证等 6 种论证要素,以及支持、质疑等论证关系。

随着科学论文内容语义增强研究的不断深入,科学论文内容本体的开发呈现出以下两种趋势:一是在科学论文内容语义建模的基础上,力求开发表达能力更强、更为全面的本体。王晓光等在功能单元理论的基础上设计并开发了一种融合情境信息的功能单元本体(Functional Units Ontology, FUO),并进行了初步的深度标注实验[53]。王晓光等也在综述论证本体的基础上,参考 DEO、DoCO 等本体进一步完善了对科学证据的定义,设计了科学论文论证本体 SAO[54]。另一方面,还有些本体更加聚焦科学论文内容的特定部分,如面向科学结论[55]、面向科学论文事件[56-57]等,力求对特定知识进行更为完备的定义。

3.2.4 科学论文细粒度内容语义组织模型

科学论文细粒度内容语义组织模型代表了对非结

构化的科学论文内容数据进行语义增强的新方向，是在将非结构化信息进行结构化、语义化之后，进行关联重组的结构性增强。

在众多内容组织模型中，最具代表性的是 T. Clark 等设计的微型出版物模型(Micropublication)[58]，该模型区分了陈述、断言、数据、方法等具备不同语义及功能的语句，并对语句间的论证关系进行了明晰。与之类似的还有 C. Bölling 等[59]提出的语义证据(Semantic Evidence,SEE)的表示方法及模型。SEE 也提供了一种以证据为线索的知识聚合方式，将特定主题的科学论断、证据与相关材料、方法、假设、推理及其他外部知识库相连接，进而形成一种相互连接且机器可读的表达。另有一些模型，以研究对象套件模型(Research Object Suit)[60]为代表，旨在提供一种结构化的容器，将研究数据与对应的研究方法以及相关的元数据封装起来，形成一个围绕特定主题的套件。

从本质上看，科学论文内容语义组织模型提供了一种新的文档表示方法。微型出版物模型和语义证据模型均把科学文献拆分成了各种论证单元，随后又根据论证结构进行了重组，既表征了科学论文内容的逻辑结构，也实现了科学论文内容组件的关联。研究对象套件模型则针对研究型论文中所包含的研究方法、实验过程与科学数据进行了关联，既表征了科学实验的过程，也为科学数据提供了较为清晰的属性及背景信息。

4　科学论文语义增强对比分析

综合以上对科学论文不同组成部分语义增强理论探索与实践的介绍，本文对不同内容数据的语义增强路径及实施情况进行了综合分析，如表 1 所示：

表 1　科学论文语义增强路径对比分析

科学论文组成部分		语义增强路径		
		语义标注	语义组织	可视化呈现
副文本内容	题录信息	BIBO、FaBIO	VIVO、Scigraph、MAG	/
	摘要	结构化摘要、SDA、亮点摘要	SDA	富媒体摘要(图像摘要、视频摘要)
	引用与参考文献	CiTO、C4O	Scigraph、MAG、OpenCitatio、AMiner	/
正文本内容	概念实体	实体抽取	/	标签云、标签树
	陈述	纳米出版物	纳米出版物	/
	内容组件	ABCDE、Discourse Segment、CISP、EXPO、AMO、DoCO、DEO、SALT、EventMine-MK、Claim Framework	微型出版物、语义证据、研究对象套件	/
	语义关系	AMO、ScholOnto、SALT	微型出版物、语义证据模型	/

由表 1 可知，对副文本内容进行语义增强的实践较为丰富。针对引用及参考文献信息，已有研究实现了对引用功能、引用情况、引用情境等的语义描述，并构建了相应的数据集。同时，借助本体及知识图谱，题录信息与参考文献信息得以关联并发布。

对正文本内容的语义增强以理论探索为主。在概念实体抽取与表示方面，借助领域本体对概念实体进行抽取与表示取得了十分突出的进展。在陈述语义描述与关联发布方面，纳米出版物数据集的建设也在稳步推进。对于内容组件、逻辑结构的语义化表示与关联，还处于理论探索阶段，虽然有相应的本体与组织模型问世，但囿于语义标注过程中的技术问题，还未能建构大规模的数据集。

总体来看，科学论文语义增强的研究与实践成果主要集中在语义描述与标注阶段，本体在语义增强过程中的重要性逐步凸显，同时也产生了诸多语义关联组织模型与知识图谱。而在内容可视化呈现方面，现有研究仍有明显不足。

5　科学论文语义增强研究的趋势研判

围绕科学论文语义增强的核心目标与关键问题，本文认为未来针对科学论文的语义增强工作及探索可以围绕以下几个方面展开：

(1)多维、多源数据的语义整合与互操作。通过不同知识图谱的设计、开发与应用，现有研究对科学论文题录信息、引用及参考文献信息进行了语义增强，但这类知识图谱较少涉及对正文本内容数据(陈述、内容组件及逻辑结构)的关联，如何在语义表征科学论文内容的基础上，填补现有科学论文知识图谱的空白，实现知识图谱与内容语义组织模型的关联，将是促进科学论文语义增强的基础。

(2)富语义内容数据的可视化。利用丰富的可视

化手段，提高科学论文内容的可感知性是科学论文语义增强的关键。目前对科学论文进行可视化的研究还十分稀少，可视化方式多以标签云、标签树等为主，可视化对象多以字词为基础。如何完整、高效、准确地可视化呈现诸如论证结构、关键信息等科学论文语义内容数据，还需要从理论与实践两个方面共同入手，进行更为深入地探索。

（3）针对科学论文领域特征的语义增强。科学论文的复杂性一方面在于其蕴含了大量知识，另一方面在于科学论文还受到了领域研究范式、研究方法及写作规范的影响。现有研究提出的内容本体、内容组织方式多面向生物医学领域，如何将现有研究成果应用到人文、社科或其他自然科学领域，还需要分领域建设表达能力更强、更为全面的内容语义表示模型，制定符合领域特征的科学论文语义增强发展方式。

（4）面向科学论文阅读行为的语义增强。科学论文语义增强的最终目标是为了帮助科研工作者快速获取论文中蕴含的大量知识，因此有必要分领域对用户的阅读任务、阅读策略、阅读对象及阅读模式进行深入研究。现阶段，有关语义增强的基础理论研究与实践探索多以对科学论文文本特征分析为起始点。无论是语义描述还是语义组织，均是建立在文本分析与逻辑分析的基础上，未能充分考虑用户的阅读特点与使用方式。因此，未来对于科学论文语义增强的研究还需要加强对用户阅读行为的理解。

6 结语

本文从语义增强概念入手，进一步规范了科学论文语义增强的概念，分析了科学论文语义增强的核心目标、实施路径与关键问题，同时在对现有理论与实践成果进行梳理的基础上进行了对比分析，总结了科学论文语义增强的特点与不足之处。总体来说，科学论文语义增强按目标的不同可以分为两类：一类是面向对科学论文内容信息的语义增强，包括论文基本信息及内容的规范化描述、论文内容实体语义标注、论文内容关联与集成等；另一类则针对科学论文内容可视化呈现，即借助多媒体实现论文内容可感知性的提升。

未来关于科学论文语义增强的研究一方面要结合领域特征，开发适用于不同领域的科学论文内容本体，创新科学论文内容多维组织方式；同时也要结合用户科学论文阅读任务、阅读策略、阅读侧重点与阅读行为模式，综合利用多种可视化方式，开发适用于用户的策略型阅读辅助工具及阅读系统，实现面向用户的科学论文语义增强。

参考文献：

[1] RENEAR A H, CAROLE L P. Strategic reading, ontologies, and the future of scientific publishing[J]. Science, 2009,325(5492): 828 – 832.
[2] SHOTTON D. Semantic publishing: the coming revolution in scientific journal publishing[J]. Learned publishing, 2009, 22(2): 85 – 94.
[3] SHOTTON D. The five stars of online journal articles: a framework for article evaluation[EB/OL]. [2020 – 12 – 20]. https//purl.pt/302/dlib/january12/shotton/olshotton.html.
[4] SHOTTON D, PORTWIN K, KLYNE G, et al. Adventures in semantic publishing: exemplar semantic enhancements of a research article[J]. PLoS computational biology, 2009, 5(4): e1000361.
[5] 翁彦琴, 李苑, 彭希珺. 英国皇家化学会(RSC)——科技期刊语义出版模式的研究[J]. 中国科技期刊研究, 2013, 24(5): 825 – 829.
[6] 翁彦琴, 彭希珺. 爱思唯尔(Elsevier)语义出版模式研究[J]. 中国科技期刊研究, 2014, 25(10):1256 – 1261.
[7] KURZ T, DAMJANOVIC V, GUNTNER G, et al. Semantic enhancement for media asset management systems[J]. Multimedia tools & applications, 2014, 70(2):949 – 975.
[8] Europeana semantic enrichment[EB/OL]. [2020 – 02 – 23]. https://pro.europeana.eu/page/europeana-semantic-enrichment.
[9] ZENG M L. Semantic enrichment for enhancing LAM data and supporting digital humanities. review article[J]. El profesional de la informacion, 2019, 28(1):1 – 35.
[10] WOUTERSEN-WINDHOUWER S, BRANDSMA R, HOOGENAAR A, et al. Enhanced publications: linking publications and research data in digital repositories[M]. Amsterdam: Amsterdam University Press, 2009.
[11] HOOGERWERF M. Durable enhanced publications[EB/OL]. [2021 – 01 – 03]. https://www.researchgate.net/publication/242732066_Durable_Enhanced_Publications.
[12] BREURE L, VOORBIJ H, HOOGERWERF M. Rich internet publications: show what you tell[J]. Journal of digital information, 2010, 12(1):1.
[13] PRASAD A R D, GIUNCHIGLIA F, DEVIKA P M. DERA: from document centric to entity centric knowledge modelling[C]// SLAVIC A, GNOLI C. Faceted classification today: theory, technology and end users: proceedings of the International UDC Seminar 2017. Würzburg: Ergon Verlag, 2017:169 – 179.
[14] Bibliographic ontology specification[EB/OL]. [2021 – 01 – 02]. http://www.bibliontology.com.
[15] PERONI S, SHOTTON D. FaBiO and CiTO: ontologies for describing bibliographic resources and citations[J]. Web semantics: science, services and agents on the World Wide Web, 2012, 17(17): 33 – 43.
[16] 张艳侠, 齐飞, 毕强. 关联数据的语义互联应用研究——以

VIVO 为实例[J]. 图书情报工作, 2013,57(17):17－21.
[17] 喻琪琛，王晓光. 科学论文摘要语义增强形式调查研究[J]. 数字图书馆论坛, 2017(8):8－15.
[18] CICCARESE P, SHOTTON D, PERONI S, et al. CiTO + SWAN: the Web semantics of bibliographic records, citations, evidence and discourse relationships[J]. Semantic Web, 2014,5(4):295－311.
[19] Citation counting and context characterization ontology [EB/OL]. [2019－12－20]. http://purl.org/spar/c4o.
[20] OpenCitation[EB/OL].[2020－12－05]. http://opencitations.net/.
[21] SciGraph [EB/OL]. [2020－05－15]. http://www.springernature.com/cn/researchers/scigraph.
[22] Aminer [EB/OL]. [2020－05－15]. https://www.aminer.cn.
[23] Microsoft academic graph [EB/OL]. [2020－05－15]. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/.
[24] Open academic graph [EB/OL]. [2020－05－15]. https://www.openacademic.ai/oag/.
[25] WANG R, YAN Y, WANG J, et al. AceKG: a large-scale knowledge graph for academic data mining[C]//Proceedings of the 27th ACM international conference on information and knowledge management. New York: Association for Computing Machinery, 2018: 1487－1491.
[26] 任海英，石彤. 科技论文微观概念地图的构建及研究思路的挖掘[J]. 图书情报工作, 2016,60(4):115－124.
[27] 丁君军,郑彦宁,化柏林. 基于规则的学术概念属性抽取[J]. 情报理论与实践, 2011,34(12):10－14,33.
[28] 乐小虬，张帆，何远标. 学术论文大纲中关键术语抽取方法研究[J]. 现代图书情报技术, 2014, 30(3): 73－79.
[29] 吴思竹，李峰，张智雄. 知识资源的语义表示和出版模式研究——以 Nanopublication 为例[J]. 中国图书馆学报, 2013(4): 102－109.
[30] KING R D, ROWLAND J, OLIVER S G,et al. The automation of science[J]. Science, 2009, 324(5923):85－89.
[31] LIAKATA M, SAHA S, DOBNIK S, et al. Automatic recognition of conceptualization zones in scientific articles and two life science applications[J]. Bioinformatics, 2012, 28(7): 991－1000.
[32] DE WAARD A, TEL G. The ABCDE format enabling semantic conference proceedings[EB/OL]. [2021－01－01]. https://www.researchgate.net/publication/220706582_The_ABCDE_Format_Enabling_Semantic_Conference_Proceedings.
[33] DE WAARD A, BUITELAAR P, EIGNER T. Identifying the epistemic value of discourse segments in biology texts[C]//Proceedings of the eighth international conference on computational semantics. Stroudsburg: Association for Computational Linguistics, 2009: 351－354.
[34] ZHANG L, KOPAK R, FREUND L, et al. A taxonomy of functional units for information use of scholarly journal articles[J]. Proceedings of the American Society for Information Science and Technology, 2010, 47(1):1－10.
[35] TEUFEL S. Argumentative zoning: information extraction from scientific text[D]. Edinburgh: University of Edinburgh, 1999.
[36] TEUFEL S. The structure of scientific articles: applications to citation indexing and summarization[M]. Stanford,CA: CSLI Publications (CSLI Studies in Computational Linguistics), 2010.
[37] GREEN N L. Representation of argumentation in text with rhetorical structure theory[J]. Argumentation, 2010, 24(2):181－196.
[38] GREEN N. Identifying argumentation schemes in genetics research articles[C]//Proceedings of the 2nd workshop on argumentation mining. Denver: Association for Computational Linguistics, 2015: 12－21.
[39] GREEN N. Argumentation mining in scientific discourse[C]//Proceedings of the 18th workshop on computational models of natural argument. London: Association for Computational Linguistics, 2017: 7－13.
[40] GREEN N. Implementing argumentation schemes as logic programs [C]//Proceedings of the 16th Workshop on computational models of natural argument. New York: Association for Computational Linguistics, 2017: 1－7.
[41] SHUM S B, MOTTA E, DOMINGUE J. ScholOnto: an ontology-based digital library server for research documents and discourse [J]. International journal on digital libraries, 2000, 3(3): 237－248.
[42] BUCKINGHAM SHUM S J, UREN V, LI G, et al. Modeling naturalistic argumentation in research literatures: representation and interaction design issues[J]. International journal of intelligent systems, 2007, 22(1): 17－47.
[43] UREN V, BUCKINGHAM SHUM S, BACHLER M, et al. Sensemaking tools for understanding research literatures: design, implementation and user evaluation[J]. International journal of human-computer studies, 2006, 64(5): 420－445.
[44] THOMPSON P, NAWAZ R, MCNAUGHT J, et al. Enriching a biomedical event corpus with meta-knowledge annotation[J]. BMC bioinformatics, 2011, 12(1). doi:10.1186/1471-2105-12-393.
[45] ANANIADOU S, THOMPSON P, NAWAZ R. Enhancing search: events and their discourse context[C]//GELBUKH A. Proceedings of the 14th international conference on computational linguistics and intelligent text processing. Berlin: Springer－Verlag, 2013: 318－334.
[46] DE WAARD A, MAAT H P. Epistemic modality and knowledge attribution in scientific discourse: a taxonomy of types and overview of features[C]//Proceedings of the workshop on detecting structure in scholarly discourse. Stroudsburg: Association for Computational Linguistics, 2012: 47－55.
[47] BLAKE C. Beyond genes, proteins, and abstracts: identifying scientific claims from full-text biomedical articles[J]. Journal of biomedical informatics, 2010, 43(2): 173－189.

[48] TUDOR G, SIEGFRIED H, KNUD M, et al. SALT – Semantically annotated LaTeX for scientific publications[C]//Proceedings of the 4th European semantic Web on the semantic Web: research and applications. Berlin: Springer-Verlag, 2007: 518 – 532.
[49] 马雨萌, 祝忠明. 科学篇章修辞块本体标准及其应用分析[J]. 情报杂志, 2012(10):112 – 116.
[50] The discourse element ontology [EB/OL]. [2020 – 05 – 15]. http://purl.org/spar/deo.
[51] CONSTANTIN A, PERONI S, PETTIFER S, et al. The document components ontology (DoCO)[J]. Semantic Web, 2016, 7(2): 167 – 181.
[52] The argument model ontology [EB/OL]. [2020 – 10 – 23]. http://www.essepuntato.it/2011/02/argumentmodel.
[53] 王晓光, 李梦琳, 宋宁远. 科学论文功能单元本体设计与标引应用实验[J]. 中国图书馆学报, 2018, 236(4):75 – 90.
[54] 王晓光, 周慧敏, 宋宁远. 科学论文论证本体设计与标注实验[J]. 情报学报, 2020, 39(9): 885 – 895.
[55] FATHALLA S, VAHDATI S, AUER S, et. al. SemSur: a core ontology for the semantic representation of research findings[C]//Proceedings of the 14th international conference on semantic systems. Vienna: Elsevier B. V., 2018: 151 – 162.
[56] JEONG S, KIM H G. SEDE: an ontology for scholarly event description[J]. Journal of information science, 2010, 36(2): 209 – 227.
[57] FATHALLA S, VAHDAT S, LANGE C, et al. SEO: a scientific events data model [EB/OL]. [2020 – 11 – 12]. https://www.researchgate.net/publication/336594094_SEO_A_Scientific_Events_Data_Model.
[58] CLARK T, CICCARESE P, GOBLE C. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications[J]. Journal of biomedical semantics, 2014, 5(1): 28.
[59] BOLLING C, WEIDLICH M, HOLZHUTTER H G. SEE: structured representation of scientific evidence in the biomedical domain using semantic Web techniques[J]. Journal of biomedical semantics, 2014, 5(S1): S1.
[60] HETTNE K M, DHARURI H, ZHAO J, et al. Structuring research methods and data with the research object model: genomics workflows as a case study[J]. Journal of biomedical semantics, 2014, 5(1):41.

作者贡献说明:
宋宁远:论文框架设计,论文撰写;
裴雷:论文思路确定,提供方向建议;
王春迎:论文修改、定稿。

The Survey and Tendency of Semantic Enrichment for Scientific Papers

Song Ningyuan[1] Pei Lei[1] Wang Chunying[2]
[1] School of Information Management, Nanjing University, Nanjing 210023
[2] School of Information Management, Zhengzhou University, Zhengzhou 450001

Abstract: [**Purpose/significance**] With the transfer of scientific communication system to electronic media, the content organization and presentation of traditional scientific papers have brought many disadvantages. Semantic enhancement of scientific papers can innovate the organization and presentation of scientific papers, which is the key to solve these problems. It has been paid attention by scientific research institutions and academic publishers and formed a series of theoretical and practical achievements. Combing and summing up these achievements and finding the advantages and disadvantages can play a guiding role in promoting the further development of semantic enhancement of scientific papers. [**Method/process**] Starting from the concept of semantic enhancement, this paper focused on the analysis of the core objectives, implementation paths and key issues of semantic enhancement in scientific papers. Then, the paper combed the theoretical and practical results of semantic enhancement of structured and unstructured data in scientific papers and made a comparative analysis by using three stages in the path of semantic enhancement of scientific papers: semantic annotation, semantic organization and visual presentation. [**Result/conclusion**] This research summarizes the characteristics of semantic enhancement of scientific papers at this stage, provides the four suggestions for the future development and research of semantic enhancement in scientific papers.

Keywords: scientific papers semantic enrichment semantic annotation semantic organization visualization